A Practical Guide to Check the Consistency of Item Response Patterns in Clinical Research Through Person-Fit Statistics

Meijer, Rob R.; Niessen, A. Susan M.; Tendeiro, Jorge N.

Published in:
Assessment

DOI:
10.1177/1073191115577800

Article

A Practical Guide to Check the Consistency of Item Response Patterns in Clinical Research Through Person-Fit Statistics: Examples and a Computer Program

Assessment
2016, Vol. 23(1) 52–62

DOI: 10.1177/1073191115577800
asm.sagepub.com
SAGE

Rob R. Meijer[1], A. Susan M. Niessen[1], and Jorge N. Tendeiro[1]

Abstract

Although there are many studies devoted to person-fit statistics to detect inconsistent item score patterns, most studies are difficult to understand for nonspecialists. The aim of this tutorial is to explain the principles of these statistics for researchers and clinicians who are interested in applying these statistics. In particular, we first explain how invalid test scores can be detected using person-fit statistics; second, we provide the reader practical examples of existing studies that used person-fit statistics to detect and to interpret inconsistent item score patterns; and third, we discuss a new *R*-package that can be used to identify and interpret inconsistent score patterns.

Keywords

validity of test scores, person fit, item response theory, data forensics

When data are collected through questionnaires or tests, respondents may produce invalid test scores as a result of, for example, lack of motivation, not understanding particular items, or faking. These types of behavior will often result in aberrant, inconsistent, or unexpected item response patterns. Therefore, in several guidelines for the quality of tests, such as in those of the International Test Commission (2013, p. 23), it is advised to "check aberrant or unexpected response patterns, (e.g., when difficult items are answered correctly and easy items incorrectly)." Also, Olson and Fremer (2013) discussed several data forensic methods to check for invalid test scores and advised to check for inconsistent response behavior.

Reynolds (2010) in his editorial for *Psychological Assessment* emphasized the importance of checking the validity of individual test scores. He stated that "reliability refers to test scores, not tests, and validity refers to the accuracy and appropriateness of *test score interpretations*—again, not to tests" (p. 3, italics added) and "each interpretation attributed to a test score must be validated: Tests are neither valid nor invalid; only the interpretations of performance on the test are the proper subject of validation research" (p. 3). Although in personality and clinical assessment the test score validity is sometimes checked for each individual test score through validity scales, in practice, checking the validity of individual test scores, be it in typical or maximum performance testing, is still not popular.

In this tutorial, we discuss the principles of person-fit statistics that are used to check the validity of test scores through the alignment between a test model and individual observed data. The test model may be a parametric item response theory (IRT) model, such as the two- or three-parameter logistic model (Embretson & Reise, 2000), or a nonparametric IRT model (Sijtsma & Molenaar, 2002), but it may also be a factor model or a set of items that are selected using classical test theory. We will not discuss the details of different statistics that are available in the literature. Instead, we give references to studies that the reader can consult for more details of individual statistics and for a more detailed overview we refer to Meijer and Sijtsma (2001). In general, the statistics we discuss in this article can be applied as long as the data are fairly unidimensional. There are many, mostly technical, papers written about these statistics by specialists (for an overview, see Meijer & Sijtsma, 2001). What is lacking is a tutorial for researchers and clinicians, that is, for those who are mainly interested in applying these statistics to analyze their data. In this article we would like to fill this gap. We first provide an empirical example that illustrates the central idea of the use of these statistics; second, we discuss the principles behind these statistics on the basis of simple group-based statistics and

[1]University of Groningen, Groningen, Netherlands

Corresponding Author:
Rob R. Meijer, Department of Psychometrics and Statistics, Faculty of Behavioral and Social Sciences, University of Groningen, Grote Kruisstraat 2/1, 9712 TS Groningen, Netherlands.
Email: r.r.meijer@rug.nl

we present existing research that illustrates how person-fit scores are related to psychological background variables, including the clinical implications in applied assessment. Finally, we discuss a recently developed *R*-package that can be used to calculate the statistics and that provide plots to diagnose the configurations of item score patterns. To be able to do this, a researcher should have access to the item scores of test takers, patients, or clients. With the popularity of computer-based assessment an increasing number of researchers will have access to such data.

We think that one of the reasons that these statistics are not often used in practice is that it is difficult for researchers to find software to calculate these statistics. Another reason is that many papers are too technical to understand for nonspecialists. Hopefully this tutorial will serve this researcher and other practitioners to apply these techniques more often.

An Illustrative Example

We start to illustrate the principles of person-fit statistics using the Physical Functioning scale of the SF-36 (Ware & Sherbourne, 1992). We use this scale because it is a well-known and often-used scale and we use it for didactic reasons: The psychometric properties of this scale are ideal to explain person fit (high discriminating items and large spread between the item difficulties). However, to further illustrate the use of person fit for psychological constructs, we will also provide results for an "Inadequacy" scale (Luteijn, van Dijk, & Barelds, 2005) at the end of this article.

In Table 1 we depicted the item content, mean item score, and examples of three item score patterns on the 10 items of the well-known Physical Functioning scale. The items are scored from 0 through 2 (0 = *limited a lot*; 1 = *limited a little*; 2 = *no, not limited at all*) and our results are based on a sample of 714 persons (see Meijer, Tendeiro, & Wanders, 2015, for a more detailed description). Meijer et al. (2015) showed that the data predominantly measured one factor and that the total score scale can be used to give an overall impression of Physical Functioning. In Table 1 the items are ordered according to decreasing mean item score in this sample, that is, to increasing amount of physical effort that is needed to fulfill the task. For each person who fills out this questionnaire, we expect that, when an item that requires a more difficult task is endorsed, such as "walking more than a mile," an "easier" item like "walking one block" is also endorsed. In practice, this will not always be the case because responses to items are probabilistic; for most persons there will be some reversals or errors (known as *Guttman errors*, i.e., answering a more "difficult" or less popular item correctly and an "easier" or more popular item incorrectly), but many errors may result in a total score that is difficult to interpret. For example, consider in Table 1 the item score patterns (items in order of increasing difficulty)

Table 1. The Physical Functioning Scale and Some Item Score Patterns.

Item	Item content	Mean	Person 493	Person 24	Person 560
PF10	Bathing or dressing yourself	1.91	1	0	0
PF9	Walking one block	1.78	1	0	0
PF3	Lifting or carrying groceries	1.78	1	2	0
PF5	Climbing one flight of stairs	1.75	1	1	0
PF8	Walking several blocks	1.68	1	2	1
PF2	Moderate activities, moving table	1.60	1	2	1
PF6	Bending, kneeling, or stooping	1.52	1	2	0
PF7	Walking more than a mile	1.43	0	1	1
PF4	Climbing several flight of stairs	1.38	0	2	2
PF1	Vigorous activities	1.00	0	2	2
G_N^p			.04	.75	.82
l_z^p			.93	−7.13	−7.94

of Person 493 [1 1 1 1 1 1 1 0 0 0] resulting in a total score of 7 and that of Person 560 [0 0 0 0 1 1 0 1 2 2], also with a total score of 7. Although both persons have the same total scores, when we consider the configuration of the item scores, the score pattern of Person 493 seems to be in line with the popularity of the items (more 0 scores on the items that require more strenuous tasks). However, it is clear that the score pattern of Person 560 is more difficult to interpret (i.e., aberrant from what is expected), because relatively easy physical tasks are difficult to conduct for this person (0 scores), whereas more difficult tasks are easier to conduct (1 and 2 scores). As a result, the total score for this person is difficult to interpret.

There may be different underlying mechanisms that cause aberrant response behavior, and below we provide examples of studies that investigated the reason of aberrant response behavior. However, irrespective of the underlying mechanisms, it is clear that the test scores of persons with unexpected answering behavior are difficult to interpret. With 714 score patterns in this sample, it is difficult to inspect every item score pattern and decide by eye-ball inspection only, which patterns are unexpected and need closer inspection. Therefore, several statistics have been proposed that provide information about how unexpected an item score pattern is. Sometimes also sampling distributions are known for these statistics under a test model, which can help a researcher to decide when to classify an item score pattern as normal or aberrant.

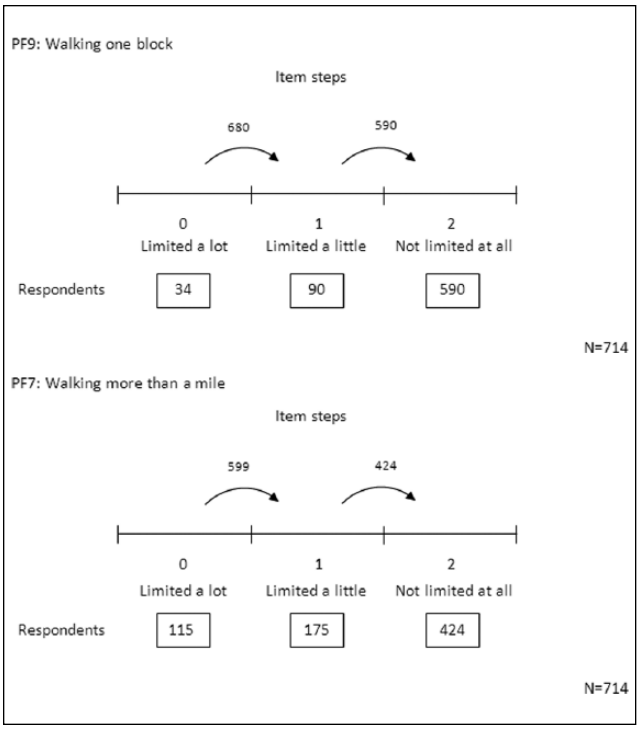

Figure 1. Illustration of item steps for two items of the PF scale.

Basic Statistics

Dichotomous Items

Almost all person-fit statistics are sensitive to the (weighted) number of Guttman errors. Consider a test consisting of 10 dichotomously scored items ordered from easy to difficult according to their item proportion-correct score. If a test taker has a total score of 6, then we expect that he or she answers the six easiest items correctly and the most difficult items incorrectly. When there is a reversal of item scores, that is, when the most difficult item is answered correctly and the easier item incorrectly, this is counted as an error. Many person-fit statistics are sensitive to the number of these errors. In general, the more errors, the more aberrant the pattern is. When we, again, assume that items are ordered from easy to difficult then a pattern [1, 1, 1, 1, 1, 1, 0, 0, 0, 0] contains zero errors, whereas the pattern [1, 1, 0, 1, 0, 1, 1, 0, 0, 1] contains 9 Guttman errors, because there are 9 (0,1) item pairs.

Polytomous Items

For polytomous items, like the items in Table 1, Guttman errors are calculated based on the number of persons that take a so-called item step. An item step is the imaginary step that a person takes when deciding to answer in a subsequent answer category. For example, consider Figure 1 where we depicted the item steps for two items "walking one block" (Item PF9) and "walking more than a mile" (Item PF7). Remember that these items have three answer categories scored 0, 1, and 2. In this case, there are two item steps: The first step is the step from 0 to 1 (and thus answering 1), and the second step is the step from 1 to 2 (answering 2). In a sample we can first determine, for each item, the number of persons that answered in each category. For Item PF9 this is: 34 persons have a 0 score, 90 persons have a 1 score, and 590 have a 2 score (see Figure 1). On the basis of these numbers we can calculate the number of persons that took the first step. For Item PF9, this is the total number of persons, 714, minus 34 persons that chose 0 = 680 (for Item PF7: 714 − 115 = 599). The number of persons that took the second step for Item PF9 equals 714 − 34 − 90 = 590 (for Item PF7: 714 − 115 − 175 = 424). As can be seen in Figure 1, there are fewer persons that took the second item step for the item "walking more than a mile" that reflects having no difficulty with a more difficult physical task than for the item "walking one block." Now, if we determine the number of persons that take each item step for all items, we can order the item steps across all items from easy to difficult. For each person, we can then compare his/her item scores with the item step ordering. Large differences point at aberrant response behavior. To further illustrate this, consider the pattern of item steps taken by person 493 (in order of increasing difficulty of the item steps): [1 1 1 1 0 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0], with 0 representing an item step not taken and 1 representing item step taken. Again, this person produces a normal response pattern with only a few Guttman errors on the item step level: There are three (0, 1) item pairs. In contrast, the pattern of item steps taken by person 24 equals [0 1 1 0 0 1 0 0 1 1 0 0 1 1 1 0 1 1]. This is an unexpected pattern with 52 Guttman errors. It is important to realize that, although the items steps are used for the calculation of the person fit statistics, for the *interpretation* of a misfitting item score pattern, the different ordering of the item steps is of not much use. Interpretation should be done on the item level. Although there are differences in the exact way the different person-fit statistics are calculated, this is the basic principle behind most statistics for polytomous item scores.

Some statistics are normed against the perfect Guttman patterns, that is, the statistic equals 0 when it perfectly fits the Guttman model and it equals 1 when it perfectly fits the reversed Guttman pattern. Thus, the higher the person-fit score, the more aberrant the item score pattern. Sometimes a statistic is normed against the maximum number of possible Guttman errors given the total score. In Table 1 we give the numerical values for the G_N^p statistic (Emons, 2008; the superscript "*p*" denotes "polytomous" and the subscript "*N*" denotes "normed"), which is normed against the maximum number of Guttman errors. Some statistics follow a standard normal distribution for a large number of

items, which eases the interpretation of the numerical values of these statistics. For example, for the l_z^p statistic (Drasgow, Levine, & Williams, 1985) values smaller than −1.96 are often interpreted as indicating unexpected response behavior at a Type I error rate of 5%. This specific asymptotic approximation does not always work as expected; an adjusted computational formula is available when the item scores are dichotomous (the so-called l_z^* statistic; Snijders, 2001; see also Magis, Raîche, & Béland, 2012, for an accessible overview). The l_z^p statistic (and its dichotomous corresponding statistic, l_z) evaluates the likelihood of a score pattern under an IRT model. Thus, to apply these statistics, it is assumed that the data can be described by an IRT model (suited to either dichotomous or polytomous items; see Embretson & Reise, 2000, for an overview of both types of models). In Table 1 we also provide values for the l_z^p statistic. As can be seen Person 493 has plausible responses, resulting in a low value for G_N^p and a positive value for l_z^p. Persons 560 and 24 provide very unlikely responses, and as a result there are extreme values for both G_N^p and for l_z^p. For example, it is strange that Person 24 has trouble climbing one flight of stairs (Item PF5) but not several flights of stairs (PF4).

Thus, person-fit statistics can be used as descriptive statistics, where misfit can be directly defined in relation to the other patterns in the sample (such as for the G_N^p statistic) or to an IRT model (such as the l_z^p statistic). In the first type of application, the question is answered: How improbable is a score pattern compared with the other persons in the sample? In the second case, the question is answered: How likely is a score pattern given the assumed IRT model? Thus, to classify an item score pattern as aberrant researchers can take, for example, the 5% or 10% most aberrant response patterns in the sample, or they may use a theoretical sampling distribution and determine in advance a Type I error rate. For most person-fit statistics the sampling distribution is unknown. Then, resampling methods can be used to obtain such a distribution (e.g., see de la Torre & Deng, 2008, for an approach based on the l_z statistic). In general, it is difficult to provide rules-of-thumb for choosing a Type I error rate. However, because falsely rejecting the null-hypothesis in most applications of person-fit research will not have large consequences (because almost always additional information will be obtained), choosing a 5% or 10% Type I error rate is a good choice.

We would like to stress that extreme person-fit scores indicate that an item score pattern is very unlikely and that, by definition, the researcher should be very careful when interpreting such total scores. The reason is that, in such cases, a person's item score pattern cannot be scaled with respect to the score patterns of other persons. Having said this, it is of course interesting to study what type of underlying mechanisms may determine aberrant response behavior. Meijer (1996) mentioned different types of possible reasons for aberrant response behavior: Sleeping (trouble getting started, and only answering items according to his/her trait value after having answered a number of items), guessing, cheating, alignment errors, plodding (working very methodologically and as a result generate almost perfect Guttman item response patterns), extremely creative response behavior, and deficiency of subabilities. However, not every type of possible aberrant response behavior could be identified empirically.

What Does an Extreme Person-Fit Score Mean?

Below, we discuss several studies that show that there is a relation between person-fit scores and different types of underlying test behavior.

Lack of Reading Skills or Interpretation Problems

Meijer, Egberink, Emons, and Sijtsma (2008) analyzed data from Harter's (1985) Self-Perception Profile for Children in a sample of children ranging from 8 through 12 years of age. They argued that, for some children, the scale scores should be interpreted with care and caution. Through combined information obtained from extreme person-fit scores, observational data, interviews, and theory about self-concept they showed that for some children item scores did not adequately reflect their trait level due to a less developed self-concept and/or problems understanding the meaning of the questions. Also, Meijer and Tendeiro (2014) found inconsistent response patterns for racial/ethnic groups in a high-stakes educational test. This was the result of insufficient knowledge of English.

Idiosyncratic Response Behavior or Traitedness

Reise and Waller (1993) and Ferrando (2012) discussed a general form of inconsistent response behavior when analyzing personality questionnaires where for some respondents the trait as measured by the items does not apply. This type of response behavior may lead to an idiosyncratic interpretation of the items, and test takers may endorse less popular items and reject easier items.

Test Taking Motivation

Ferrando (2012) found a group of students filling out a Neuroticism and Extraversion scale (for research purposes) that were low in what he called "person reliability," that is, the test-takers were largely insensitive to the ordering of the items. As a consequence, the resulting response patterns were almost random. This was probably due to unmotivated test behavior. In a recent study, Niessen, Meijer, and Tendeiro (2014) found that when collecting data for research

purposes some students' response patterns were also more or less random as a result of unmotivated test behavior, resulting in extreme person-fit scores. Also Schmitt, Cortina, and Whitney (1993) showed using hierarchical regression analysis that person-fit scores added information to criterion scores in a validation sample. When they removed persons with extreme person-fit scores validities increased.

Personality and Pathology

Conijn (2013) investigated consistency of response behavior on the Outcome Questionnaire-45 and found that patients having more severe distress and patients with psychotic disorders, somatoform disorders, and substance-related disorders were particularly likely to show misfit. Conijn (2013) concluded that person-fit analysis has potential in outcome measurement for subgroups that are at risk of producing invalid results. Conrad et al. (2010) screened for atypical suicide risk for persons enduring alcohol and drug treatment. They found that extreme person-fit scores indicated patients that endorsed suicidal ideation items, whereas these persons did not endorse items that reflect less severe forms of depressed feelings.

Other researchers like Meijer and van Krimpen-Stoop (2001) and Woods, Oltmanns, and Turkheimer (2008) found that males produced more inconsistent responses than females. Schmitt, Chan, Sacco, McFarland, and Jennings (1999) and Ferrando (2009) found that respondents low on conscientiousness generated less consistent response patterns than test takers high in conscientiousness.

Zickar and Drasgow (1996) analyzed data in which respondents were instructed to answer honestly or to fake the good answer on a personality scale. They found that using the person-fit approach a larger number of faking respondents were detected at low false positive rates than using a social desirable scale.

Recently, Wardenaar, Wanders, Roest, Meijer, and de Jonge (2014) investigated the usefulness of person-fit statistics to answer the question whether observed association between depression and acute coronary syndromes are overestimated due patients' tendency to endorse mainly somatic items on depression scales. Wardenaar et al. (2014) using the Beck Depression Inventory (BDI) data found that somatic items were most often endorsed by the majority of persons and that atypical patients scored higher on the depressive mood/cognitions (e.g., "sense of failure") and lower on somatic items. The clinical implication is that person-fit statistics do tell us that (a) relatively high scores on the BDI are due to high scores on somatic complaints and (b) BDI scores from atypical respondents should be interpreted differently than BDI scores from typical respondents. That is, for typical respondents the association between depression and acute coronary syndromes seems to be overestimated, whereas for atypical respondents it is not. Accounting for interpersonal differences in item responding can help to improve the validity of depression assessments in psychosomatic research.

Table 2. Person-Fit Statistics Available in the *R* Package PerFit.

		Type of data	
Statistics	Reference	Dichotomous	Polytomous
Nonparametric			
r.pbis	Donlon and Fischer (1968)	x	
C	Sato (1975)	x	
G, G_n	van der Flier (1980); Meijer (1994)	x	
A, D, E	Kane and Brennan (1980)	x	
U3, ZU3	van der Flier (1982)	x	
*C**	Harnisch and Linn (1981)	x	
NCI	Tatsuoka and Tatsuoka (1982, 1983)	x	
H^t	Sijtsma (1986)	x	
G^p	Molenaar (1991)		x
G_N^p	Molenaar (1991), Emons (2008)		x
$U3^p$	Emons (2008)		x
Parametric			
l_z	Drasgow et al. (1985)	x	
l_z^p	Drasgow et al. (1985)		x
l_z^*	Snijders (2001)	x	

Which Person-Fit Statistic to Use?

There are numerous person-fit statistics. In Meijer and Sijtsma (2001), an overview is given of different statistics. There are also studies in which the power of different statistics is compared using simulated data. Karabatsos (2003) concluded that the H^t coefficient had the highest power given a controlled false positive rate. An interesting conclusion from his study was that group-based statistics like H^t, *C*, and *U3* (for references, see Table 2) outperformed parametric statistics like l_z. A researcher can calculate these statistics using the program PerFit (Tendeiro, 2015) described below. Also, Tendeiro and Meijer (2014) recently compared different group-based statistics for dichotomous item scores. They concluded that, given a fixed Type I error rate of .05 in general H^t, followed by *U3*, and *C* had the highest power to detect misfitting response vectors.

Another criterion to select a person-fit statistic is whether there is a theoretical sampling distribution available. For

dichotomous data and parametric IRT analyses, the l_z^* statistic is to be recommended because of the availability of clearly interpretable cutoff scores, that is, for l_z^* its distribution has a better approximation to the assumed standard normal distribution than the l_z statistic. Thus, for l_z^* the Type I error rate is easier to control than for other statistics. For polytomous items and long tests also l_z is most often used, probably because of the easy interpretation.

Drasgow, Levine, and McLaughlin (1991) proposed a multiple-subtest extension of l_z denoted l_{zm}. Recently, Conijn (2013) compared the use of l_z in combination with l_{zm} and found that for noncognitive data a combination of both statistics resulted in the highest power for a fixed Type I error rate.

Strategy to Detect and Interpret Aberrant Response Patterns

Rupp (2013), following Meijer and Sijtsma (2001), recently formulated some guidelines on how to perform a person-fit analysis. He suggested that each person-fit analysis should contain "(1) a statistical detection step," where the scores on at least one person-fit statistic are computed; "(2) a numerical tabulation step," where the item score patterns are shown; "(3) a graphical exploration step" to display the item response patterns; "(4) a quantitative exploration step" that takes into account possible covariates; and "(5) a qualitative explanation step," such as think-aloud protocols or interviews that may help to explain aberrant response behavior. Using the *R*-package discussed below, Steps 1 through 3 can be performed easily. With respect to Steps 4 and 5, it depends on the type of application envisaged by the researcher whether these steps should and can be taken. In Step 4 scores on other tests can be incorporated in the analysis. As we discussed above, Ferrando (2009), for example, found that persons low on conscientiousness were more likely to produce inconsistent response patterns than persons scoring high on conscientiousness; other researchers found differences between males and females (e.g., Meijer & van Krimpen-Stoop, 2001). In Step 5, a clinician may be able to discuss extreme person-fit scores (aberrant response behavior) with a client or with other informants. A clinical interview may provide such a setting. However, Steps 4 and 5 are sometimes difficult to realize or may not always be of interest. For example, removing very unlikely item score patterns from unmotivated students that fill out a questionnaire for research purposes may be appropriate without giving an additional explanation. However, as Meijer and Tendeiro (2014) pointed out, in high-stakes educational testing it is not always easy to withhold test-takers a test score on the basis of person-fit scores (or any other "statistical evidence"). In these contexts, however, it may help understand particular types of scores. For example, as discussed above, Meijer and Tendeiro (2014) found that several test takers with language problems produced unexpected response patterns as a result of guessing and thus scored low on the test.

How Should We Manage Unexpected Score Patterns and Clinical Implications of Person-Fit Scores?

An individual's total score that is based on very unlikely item scores is difficult to interpret. Person-fit statistics thus function as a validity index for individual test scores. How to proceed with a person who produces an unlikely score pattern? This depends on the test setting. In an interesting article, Conrad et al. (2010) showed that through the use of person-fit statistics, persons could be identified with an atypical subtype at high risk for suicide that did not show typical depression and other internalizing disorders. Clinically, this is interesting information that adds additional information to the total score and that may help clinicians become more sensitive to the additional information that can be retrieved from a total score. In this case, additional information can be obtained in a clinical interview. Also, the study by Wardenaar et al. (2014) showed that BDI scores should be interpreted carefully and that total scores have different meanings for different persons. As another example, with the increasing use of computer-based administration of clinical questionnaires through small electronic devices (e.g., phones) checking the validity of test scores becomes even more important. Patients are asked to fill out a questionnaire at irregular intervals and motivation problems may arise. In these cases, it is important to check the validity of test scores and patients can be asked to rethink answers or to provide additional responses (retesting) when responses are very inconsistent.

A Computer Program and Examples

Computer Program. To help researchers calculate person-fit statistics and to interpret misfitting item score patterns, the *R*-package PerFit (Tendeiro, 2015; Tendeiro, Meijer, & Niessen, 2015) can be used. Although there are some programs that can be used to calculate person-fit statistics, such as WPERFIT (Ferrando & Lorenzo, 2000) and PERSON (Choi, 2010), one advantage of PerFit is that the program only needs a data file with item scores and that item and person parameters can be estimated through the program. However, the user can also use already estimated item parameters. Another advantage of PerFit is that the package calculates a number of group-based statistics that are not available in other packages such as, for example, the powerful H^T statistic. In Table 2, we provide an overview of all the statistics that can currently be calculated using the program

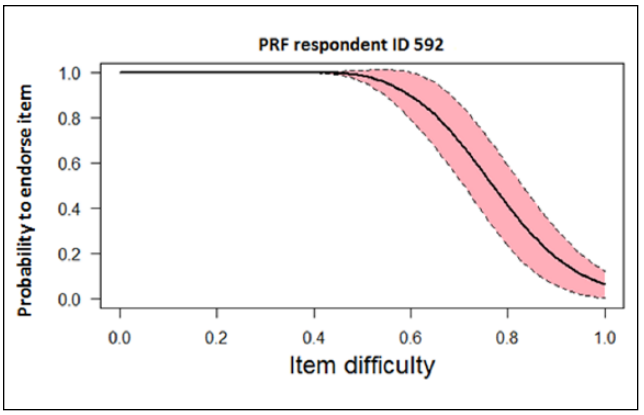

Figure 2. Person response function for a normal behaving person on the Inadequacy scale.

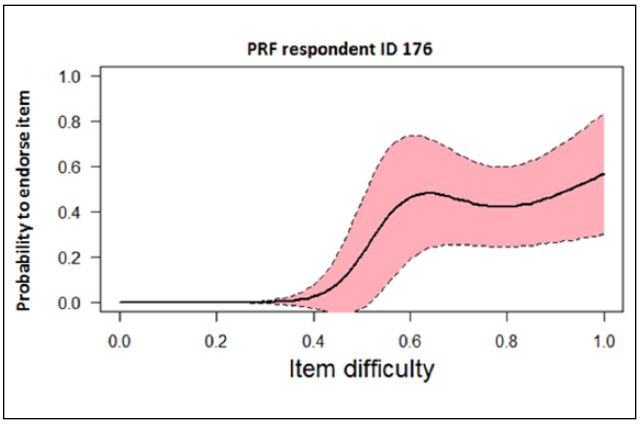

Figure 3. Person response function for an aberrant respondent on the Inadequacy scale.

(more methods will be added in the near future). There are also references to articles that discuss the individual statistics. Examples of syntax for computing the statistics and for creating plots and cutoff scores are in the help files of the package. Furthermore, the program plots nonparametric person response functions (PRFs; Emons, Sijtsma, & Meijer, 2004; Sijtsma & Meijer, 2001) for dichotomous item score patterns. PRFs give a visual representation of the fit of an individual item score pattern. More specifically, a PRF gives the probability of answering an item correctly or endorsing an item as a function of the item difficulty. For example, in Figures 2 and 3 we depict two examples of PRFs for persons who completed a dichotomously scored Inadequacy scale (to be discussed in more detail below). The PRFs for Person 592 (normal response behavior; G_N = .04) and Person 176 (aberrant response behavior; G_N = .60) are shown. The *y*-axis represents the probability of endorsing an item and the *x*-axis displays the difficulty values (1 − proportion-correct score), which is an indicator of item popularity (higher values indicating less popular items). Person 592 (see Figure 2) displays an expected pattern: The less popular the items are, the smaller the probability that this person endorses an item. The PRF of Person 176 (see Figure 3), on the contrary, seems to show that for less popular items, the probability of endorsing becomes larger. This is, of course, strange. Such response behavior is worth further inspection for a full interpretation of the results.

Examples. To further illustrate the use of PerFit, we discuss two empirical examples. The first example is concerned with the detection of unmotivated students in the context of a web survey. Data were collected from 294 students for research purposes in a population of students at a university. The survey consisted of several short questionnaires, including the Big Five Inventory (John, Donahue, & Kentle, 1991; John, Naumann, & Soto, 2008), a medium length personality questionnaire. The students had to participate in a number of studies in order to receive course credit. Because the students had no direct gain in filling out the questionnaire seriously, there were concerns about the motivation of the students to complete the questionnaires and thus with respect to general data quality. We used person-fit analysis to detect careless responding students. One of the statistics used was the normed number of Guttman errors, which was computed for each subscale using PerFit. Respondents with aberrant responses could be identified and removed from the data set after inspection. An example is a female student 113, who had normed Guttman errors in the top 5% on two subscales, Extraversion (G_N^P = .37) and Neuroticism (G_N^P = .41). Inspecting the score pattern showed that she strongly endorsed the statement that she can be tense, but also agreed that she is relaxed and handles stress well. To further validate this extreme person-fit score, we inspected some other indicators of aberrant response behavior (see Meade & Craig, 2012). Further inspection showed that this student also responded suspiciously to an instructed response bogus item and completed the questionnaire in 14 minutes, whereas for most students the estimated completion time was about 20 minutes. Based on this information it can be defended to remove this item score pattern from the data.

As another example, we analyzed data from the Inadequacy scale from the Dutch Personality Questionnaire Junior (Dutch: Nederlandse Persoonlijkheidsvragenlijst-Junior: NPV-J; Luteijn et al., 2005). The NPV-J is a Dutch personality scale for adolescents and consists of five subscales. It is based on the California Personality Inventory. The Inadequacy scale was selected because it has good psychometric properties (Weekers & Meijer, 2008) and a relatively large number of items. Data were collected from 866 persons who attended primary and secondary education. Each item was either answered with "agree" or "disagree." Figure 4 shows the distribution of G_N. Note that, in this

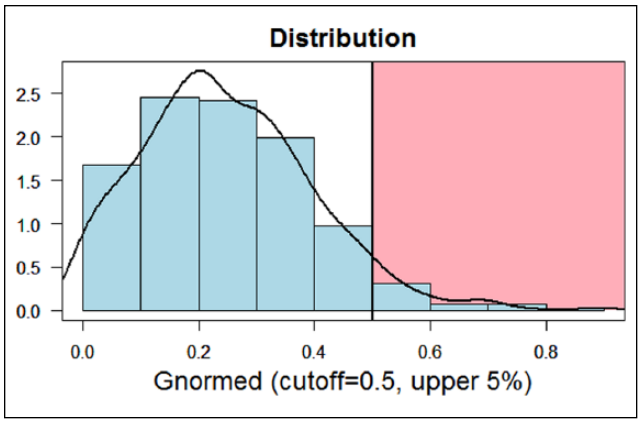

Figure 4. Histogram of the sample values of the G_N statistic of the Inadequacy scale of the NPV-J, with a 5% cutoff score.

sample, G_N values larger than .5 belong to the 5% most extreme patterns. Table 3 provides the item content ordered in decreasing popularity and three item score patterns with the corresponding G_N, H^t, and l_z^* values. The item score pattern of Person 592 has a good fit. The respondent hardly endorsed unpopular items without endorsing more popular items. In contrast, the score pattern of Person 176 is an example of an aberrant response pattern. Popular items are hardly endorsed, whereas less popular items are endorsed. Although item content is more ambiguous than for the Physical Functioning scale, it is interesting to consider the different item scores. Content-wise it is strange that this person endorses Item IN91 ("I start to sweat when I think of difficult homework"), but did not endorse Item IN98 ("When I fail at something I tried my best for I feel sad"). Also, this person endorses Item IN28 ("hate myself"), without endorsing milder signs of Inadequate feelings. The response pattern of Person 1,066 is more difficult to interpret. The endorsed items are about evenly spread out over the difficulty ordering, which may indicate unmotivated test behavior and as a result random response behavior.

Limitations

Person-fit statistics are sensitive to unexpected response patterns conditional on the total score or trait value. This implies that if a respondent decides to fill out a questionnaire very consistently, for example, only choosing the first option resulting in, for example, all 1 scores, this pattern is not necessarily unexpected given the total score or trait value and such a pattern will not be detected as misfitting. Thus, person-fit statistics are especially sensitive to *inconsistent* response patterns. When researchers are interested in detection of, for example, long strings of similar scores that may be the result of unmotivated answering behavior, validity scales may be used or special person-fit statistics that are aimed at detecting long strings of unexpected responses (see Tendeiro & Meijer, 2012). Another limitation is that the power of person-fit statistics depends on test length and item

Table 3. Item score patterns and person-fit statistics for the Inadequacy scale.

Item	Item content	Mean	Person 592	Person 176	Person 1,066
IN98	When I fail I feel sad	.56	1	0	1
IN100	Nobody can do anything right	.50	1	0	0
IN52	When I am working on something I often think of other things	.40	1	0	0
IN34	Tired often	.33	1	0	0
IN70	Others talk about me	.33	1	1	0
IN48	Dream about things that I would not tell others	.32	1	1	1
IN50	Ponder often	.30	1	1	0
IN57	Nervous often	.28	1	1	0
IN32	Worry about what I look like	.26	0	0	0
IN4	Often get headache when I worry	.24	1	0	0
IN96	Feel like something bad is going to happen	.24	1	0	0
IN102	Others are bothered when I am around	.24	1	0	0
IN6	Often bad mood	.20	1	0	1
IN36	Few people understand me	.19	1	1	0
IN72	Often wake up at night	.19	0	1	0
IN93	Often fail	.19	0	1	0
IN19	When excited, my voice sounds strange	.17	0	0	0
IN59	Always get blamed when things go wrong	.17	1	0	0
IN13	Scared in the dark	.15	0	0	0
IN75	Often bad dreams	.15	0	1	0
IN14	Others are happier than me	.14	0	0	0
IN1	Cry for no reason	.13	0	0	1
IN8	Feel worthless	.11	0	0	0
IN91	Sweat when thinking of difficult homework	.11	0	1	0
IN38	Nobody loves me	.09	0	1	1
IN66	I wish I wasn't born	.09	0	0	0
IN54	Sad often	.08	0	1	1
IN28	Hate myself	.06	0	1	0
G_N			.04	.60	.55
l_z^*			1.96	−2.15	−2.70
H^t			.43	−.13	.01

characteristics. Longer tests will result in higher power to detect aberrant response behavior than shorter tests. In general, for short scales, say less than 10 items, it is not advisable to use person-fit statistics. Also, high item discrimination and large spread between the difficulties of the items will result in high power. Fortunately, however, many tests that consist of items with low discrimination power are often long (educational tests), whereas tests that are relatively short consists of items with relatively high discriminating items (stand-alone clinical scales, such as the Beck Depression Inventory).

Discussion

As several authors have emphasized, it is important to check for the validity of individual test scores. This can be done using different methods; in this article, we discussed the use of person-fit statistics. Of course, no method or even combination of methods can be used as final *evidence* that irregular activities have taken place. Nevertheless, we believe that ignoring the information that these methods can provide is like a detective "ignoring the footprints in the garden because they might not belong to the burglar" (Funder, 2007, p. 22). When psychological and educational tests are used for individual decision making, it is important to check for the individual validity of test scores. The methods and the software package that we discussed in this study can be of great help for researchers and practitioners.

In the literature, there is some discussion about the "validity of validity indices." For example, McGrath, Mitchel, Kim, and Hough (2010) did not find much evidence for the influence of individual response configurations ("response styles") on the validity of test scores, whereas others (e.g., Holden, Wheeler, & Marjanovic, 2012; Sotaridona & Meijer, 2003) showed that it is important to investigate these configurations of item scores. An important distinction between authors in favor or against the use of validity indices lies in the way these effects are evaluated. McGrath et al. (2010) used moderated multiple regression analysis in evaluating the presence of an effect of aberrant responding. In their study, criterion scores are regressed onto predictor scores (clinical scale scores, like BDI scores) and moderator scores (e.g., scores on the validity scale or person-fit scores) and their product term. The statistical significance of this product term then determines whether there is a moderator effect of validity scores. Above we discussed the study by Schmitt et al. (1993) who showed that the product of person-fit scores and total scores is significant. However, even if this product is not significant, several authors have argued (e.g., Ben-Porath & Waller, 1992; Holden et al., 2012) that aberrant responding can still influence the validity of *individual* test scores.

Also note that McGrath et al. (2010) "focus exclusively on studies that evaluated whether response bias indicator suppress or moderate the validity of substantive indicators" (p. 454). They found that "the evidence was simply insufficient to draw firm conclusions" (p.436) or "the evidence failed to corroborate the hypothesis, although careless responding represents a possible exception" (p. 463). However, their final conclusion was rather nuanced saying that "in the absence for or against the measurement of response bias, a reasonable argument can be made for the continuing to use response bias indicators," but they warned against the detrimental effects of false positives, that is, erroneously concluding that a test taker has cheated whereas in practice this is not the case. Especially, when scores are corrected on the basis of bias indicators this may affect the rank ordering in, for example, personnel selection contexts.

We fully agree with their remark and we are certainly not in favor of correcting scores on the basis of, for example, person-fit scores. At the same time, their remark shows that manipulating responses may result in different ordering of persons and may have an important effect on selection results. Also note that person-fit statistics are especially useful to detect careless or random responding. Furthermore, Holden et al. (2012) showed that although moderated multiple regression may indicate that criterion validity can vary as a function of another variable, this technique does "not identify specific cases where another variable has manifested its effect" (p. 19). They showed that discriminant analysis was better able to identify biases responses than moderated multiple regression techniques.

On a more general level, ignoring information about the validity of individual test scores because on a group level there seems to be no effect between predictor scores and validity indices is not to be recommended. For example, in high-stakes educational testing, individual test scores are checked and students with very unlikely score patterns or very similar score patterns as other students have to redo the exam (Belov, personal communication, April 4, 2013). Note that in this example it is important to check the total scores, irrespective of the effect on the group level. Also in clinical research, several studies discussed above showed that total scores for persons with extreme person-fit scores are difficult to compare. For example, in the study by Wardenaar et al. (2014), clients with extreme person-fit scores endorse items that refer to mood/cognition and do not endorse the somatic items. Depression scores for persons with extreme person-fit scores should thus be interpreted differently than total scores for persons with normal person-fit scores.

Finally, note that person-fit statistics define invalid scores with respect to "group-level" patterns. In person-fit research, it is *always* assumed that for a large majority of persons the test model can be used as a valid indicator of their trait level. However, for a small percentage the response patterns may be very unlikely given the IRT model and further inspection may be required. This is also in agreement with Dawes, Faust, and Meehl's (1989) observation that in general it would be absurd to rely on an actuarial model (i.e., in our

case an IRT model) if there is overwhelming evidence that this model will not apply because of particular observations. This is known as the "broken leg problem" where an

> actuarial formula can be successfully used to predict an individual's attendance to the movie is highly successful in predicting an individual's weekly attendance at a movie but should be discarded upon discovering that the subject is in a cast with a fractured femur. The clinician may beat the actuarial method if able to detect the rare fact and decide accordingly. (Dawes et al., 1989, p.1670)

As discussed above, person-fit statistics may help identify such rare cases.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

Ben-Porath, Y. S., & Waller, N. G. (1992). Five big issues in clinical personality assessment: A rejoinder to Costa and McCrae. *Psychological Assessment*, *4*, 23-25.

Choi, S. W. (2010). PERSONz: Person misfit detection using the lz statistic and Monte Carlo simulations. *Applied Psychological Measurement*, *34*, 457-458.

Conijn, J. M. (2013). *Detecting and explaining person misfit in non-cognitive measurement* (Unpublished doctoral dissertation). Tilburg University, Netherlands.

Conrad, K. J., Bezruczko, N., Chan, Y., Riley, B., Diamond, G., & Dennis, M. L. (2010). Screening for atypical suicide risk with person fit statistics among people presenting to alcohol and other drug treatment. *Drug and Alcohol Dependence*, *106*, 92-100. doi:10.1016/j.drugalcdep.2009.07.023

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science*, *243*, 1668-1674.

de la Torre, J., & Deng, W. (2008). Improving person fit assessment by correcting the ability estimate and its reference distribution. *Journal of Educational Measurement*, *45*, 159-177.

Donlon, T. F., & Fischer, F. E. (1968). An index of an individual's agreement with group-determined item difficulties. *Educational and Psychological Measurement*, *28*, 105-113. doi:10.1177/001316446802800110

Drasgow, F., Levine, M. V., & McLaughlin, M. E. (1991). Appropriateness measurement for some multidimensional test batteries. *Applied Psychological Measurement*, *15*, 171-191. doi:10.1177/014662169101500207

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*, *38*, 67-86. doi:10.1111/j.2044-8317.1985.tb00817.x

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Emons, W. M. (2008). Nonparametric person-fit analysis of polytomous item scores. *Applied Psychological Measurement*, *32*, 224-247. doi:10.1177/0146621607302479

Emons, W. M., Sijtsma, K., & Meijer, R. R. (2004). Testing hypotheses about the person-response function in person-fit analysis. *Multivariate Behavioral Research*, *39*, 1-35. doi:10.1207/s15327906mbr3901_1

Ferrando, P. (2009). A graded response model for measuring person reliability. *British Journal of Mathematical and Statistical Psychology*, *62*, 641-662. doi:10.1348/000711008X377745

Ferrando, P. J. (2012). Assessing inconsistent responding in E and N measures: An application of person-fit analysis in personality. *Personality and Individual Differences*, *52*, 718-722. doi:10.1016/j.paid.2011.12.036

Ferrando, P. J., & Lorenzo, U. (2000). WPerfit: A program for computing parametric person-fit statistics and plotting person response curves. *Educational and Psychological Measurement*, *60*, 479-487.

Funder, D. (2007). *The personality puzzle*. New York, NY: W. W. Norton.

Harnisch, D. L., & Linn, R. L. (1981). Analysis of item response patterns: Questionable test data and dissimilar curriculum practices. *Journal of Educational Measurement*, *18*, 133-146. doi:10.1111/j.17453984.1981.tb00848.x

Harter, S. (1985). *Manual for the self-perception profile for children*. Denver, CO: University of Denver.

Holden, R.C., Wheeler, S., & Marjanovic, Z. (2012). When does random responding distort self-report personality assessment? An example with the NEO PI-R. *Personality and Individual Differences*, *52*, 15-20.

International Test Commission. (2013). *ITC guidelines on quality control in scoring, test analysis, and reporting of test scores*. Retrieved from http://www.intestcom.org/Guidelines/Quality+Control.php

John, O. P., Donahue, E. M., & Kentle, R. L. (1991). *The Big Five Inventory Versions 4a and 54*. Berkeley: University of California, Berkeley, Institute of Personality and Social Research.

John, O. P., Naumann, L. P., & Soto, C. J. (2008). Paradigm shift to the integrative Big Five trait taxonomy: History, measurement, and conceptual issues. In O. P. John, R. W. Robins & L. A. Pervin (Eds.), *Handbook of personality: Theory and research* (pp. 114-158). New York, NY: Guilford Press.

Kane, M. T., & Brennan, R. L. (1980). Agreement coefficients as indices of dependability for domain-referenced tests. *Applied Psychological Measurement*, *4*, 105-126. doi:10.1177/014662168000400111

Karabatsos, G. (2003). Comparing the aberrant response detection performance of Thirty-Six Person-Fit Statistics. *Applied Measurement In Education*, *16*, 277-298. doi:10.1207/S15324818AME1604_2

Luteijn, F., van Dijk, H., & Barelds, D. P. H. (2005). *NPV-J: Junior Nederlandse Persoonlijkheidsvragenlijst. Herziene handleiding 2005* [NPV-J: Dutch Personality Questionnaire–Junior: Professional manual (revised)]. Amsterdam, Netherlands: Harcourt Assessments.

Magis, D., Raîche, G., & Béland, S. (2012). A didactic presentation of Snijders's lz* index of person fit with emphasis on response model selection and ability estimation. *Journal of Educational and Behavioral Statistics*, *37*, 57-81. doi:10.3102/1076998610396894

McGrath, M. E., Mitchel, M., Kim, B. H., & Hough, L. (2010). Evidence for response bias as a source of error variance in applied measurement. *Psychological Bulletin*, *136*, 450-470.

Meade, A. W., & Craig, S. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*, 437-455. doi:10.1037/a0028085

Meijer, R. R. (1994). The number of Guttman errors as a simple and powerful person-fit statistic. *Applied Psychological Measurement*, *18*, 311-314. doi:10.1177/014662169401800402

Meijer, R. R. (1996). Person-fit research: An introduction. *Applied Measurement in Education*, *9*, 3-8.

Meijer, R. R., Egberink, I. J. L., Emons, W. H. M., & Sijtsma, K. (2008). Detection and validation of unscalable item score patterns using item response theory: An illustration with Harter's self-perception profile for children. *Journal of Personality Assessment*, *90*, 227-238. doi:10.1080/00223890701884921

Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement*, *25*, 107-135. doi:10.1177/01466210122031957

Meijer, R. R., & Tendeiro, J. N. (2014). *The use of person-fit scores in high stakes educational testing: How to use them and what they tell us* (LSAC Research Report 14-03). Retrieved from http://www.lsac.org/lsacresources/research/all/rr

Meijer, R. R., Tendeiro, J. N., & Wanders, R. B. K. (2015). The use of nonparametric item response theory to explore data quality. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment* (pp. 85-110) . London, England: Routledge.

Meijer, R. R., & van Krimpen-Stoop, E. W. L. A. (2001). Person fit across subgroups: An achievement testing example. In A. Boomsma, M. A. J. van Duijn & T. A. B. Snijders (Eds.), *Essays on item response theory* (pp. 377-390). New York, NY: Springer-Verlag.

Molenaar, I. W. (1991). A weighted Loevinger H-coefficient extending Mokken scaling to multicategory items. *Kwantitatieve Methoden*, *12*, 97-117.

Niessen, A. S. M., Meijer, R. R., & Tendeiro, J. N. (2014). *Detection of unmotivated students in a realistic test setting*. Manuscript in preparation.

Olson, J. F., & Fremer, J. (2013). *TILSA Test security guidebook: Preventing, detecting, and investigating test security irregularities*. Washington, DC: Council of Chief State School Officers.

Reise, S. P., & Waller, N. G. (1993). Traitedness and the assessment of response pattern scalability. *Journal of Personality and Social Psychology*, *65*, 143-151. doi:10.1037/0022-3514.65.1.143

Reynolds, C. R. (2010). Measurement and assessment: An editorial view. *Psychological Assessment*, *22*, 1-4. doi:10.1037/a0018811

Rupp, A. A. (2013). A systematic review of the methodology for person fit research in item response theory: Lessons about generalizability of inferences from the design of simulation studies. *Psychological Test and Assessment Modeling*, *55*, 3-38.

Sato, T. (1975). *The construction and interpretation of S-P tables*. Tokyo, Japan: Meishi Tosho.

Schmitt, N., Chan, D., Sacco, J. M., McFarland, L. A., & Jennings, D. (1999). Correlates of person fit and effect of person fit on test validity. *Applied Psychological Measurement*, *23*, 41-53. doi:10.1177/01466219922031176

Schmitt, N., Cortina, J. M., & Whitney, D. J. (1993). Appropriateness fit and criterion-related validity. *Applied Psychological Measurement*, *17*, 143-150.

Sijtsma, K. (1986). A coefficient of deviance of response patterns. *Kwantitative Methoden*, *7*, 131-145.

Sijtsma, K., & Meijer, R. R. (2001). The person response function as a tool in person-fit research. *Psychometrika*, *66*, 191-207. doi:10.1007/BF02294835

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory*. Thousand Oaks, CA: Sage.

Snijders, T. B. (2001). Asymptotic null distribution of person fit statistics with estimated person parameter. *Psychometrika*, *66*, 331-342. doi:10.1007/BF02294437

Sotaridona, L. S., & Meijer, R. R. (2003). Two new statistics to detect answer copying. *Journal of Educational Measurement*, *40*, 53-69.

Tatsuoka, K. K., & Tatsuoka, M. M. (1982). Detection of aberrant response patterns and their effect on dimensionality. *Journal of Educational Statistics*, *7*, 215-231. doi:10.2307/1164646

Tatsuoka, K. K., & Tatsuoka, M. M. (1983). Spotting erroneous rules of operation by the individual consistency index. *Journal of Educational Measurement*, *20*, 221-230. doi:10.1111/j.1745-3984.1983.tb00201.x

Tendeiro, J. N. (2015). PerFit (version 1.3) [Computer software]. University of Groningen. Retrieved from http://r-forge.r-project.org/R/?group_id=1878

Tendeiro, J. N., & Meijer, R. R. (2012). A CUSUM to detect person misfit: A discussion and some alternatives for existing procedures. *Applied Psychological Measurement*, *36*, 420-442. doi:10.1177/0146621612446305

Tendeiro, J. N., & Meijer, R. R. (2014). Detection of invalid test Scores: The usefulness of simple nonparametric statistics. *Journal of Educational Measurement*, *51*, 239-259. doi:10.1111/jedm.12046

Tendeiro, J. N., Meijer, R.R., & Niessen, A. S. M. (2015). PerFit: An R Package for person fit in IRT. Manuscript submitted for publication.

van der Flier, H. (1980). *Vergelijkbaarheid van individuele testprestaties* [Comparability of individual test performance]. Lisse, Netherlands.

van der Flier, H. (1982). Deviant response patterns and comparability of test scores. *Journal of Cross-Cultural Psychology*, *13*, 267-298. doi:10.1177/0022002182013003001

Wardenaar, K. J., Wanders, R. B. K., Roest, A. M., Meijer, R. R., & de Jonge, P. (2014). *What do depression questionnaires measure in patients with acute coronary syndromes? A psychometric investigation using item response theory and person fit*. Manuscript submitted for publication.

Ware, J. E., Jr., & Sherbourne, C. D. (1992). The MOS 36-item short-form health survey (SF-36): Conceptual framework and item selection. *Medical Care*, *30*, 473-483.

Weekers, A. M., & Meijer, R. R. (2008). Scaling response processes on personality items using unfolding and dominance models: An illustration with a Dutch dominance and unfolding personality inventory. *European Journal of Psychological Assessment*, *24*, 65-77.

Woods, C. M., Oltmanns, T. F., & Turkheimer, E. (2008). Detection of aberrant responding on a personality scale in a military sample: An application of evaluating person fit with two-level logistic regression. *Psychological Assessment*, *20*, 159-168. doi:10.1037/1040-3590.20.2.159

Zickar, M. J., & Drasgow, F. (1996). Detecting faking on a personality instrument using appropriateness measurement. *Applied Psychological Measurement*, *20*, 71-87.